World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:9, No:08, 2015

New Two-Way Map-Reduce Join Algorithm: Hash Semi Join

Authors : Marwa Hussein Mohamed, Mohamed Helmy Khafagy, Samah Ahmed Senbel

Abstract : Map Reduce is a programming model used to handle and support massive data sets. Rapidly increasing in data size and big data are the most important issue today to make an analysis of this data. map reduce is used to analyze data and get more helpful information by using two simple functions map and reduce it's only written by the programmer, and it includes load balancing , fault tolerance and high scalability. The most important operation in data analysis are join, but map reduce is not directly support join. This paper explains two-way map-reduce join algorithm, semi-join and per split semi-join, and proposes new algorithm hash semi-join that used hash table to increase performance by eliminating unused records as early as possible and apply join using hash table rather than using map function to match join key with other data table in the second phase but using hash tables isn't affecting on memory size because we only save matched records from the second table only. Our experimental result shows that using a hash table with hash semi-join algorithm has higher performance than two other algorithms while increasing the data size from 10 million records to 500 million and running time are increased according to the size of joined records between two tables.

Keywords : map reduce, hadoop, semi join, two way join
Conference Title : ICCTA 2015 : International Conference on Computing : Theory and Applications
Conference Location : London, United Kingdom
Conference Dates : August 20-21, 2015